Gu Test: A Progressive Measurement of Generic Artificial Intelligence

Author: Lifan Gu□□□□□

1. Introduction

Artificial intelligence (AI) is very different from traditional computer systems. Traditional hardware is designed with clear logic and capacity. Traditional software is also designed with clear logic goals usually: what it should be able to do and what it should not be able to do.

However, many AI technologies are very unstable, with empirical learning processes and empirical parameter tuning, etc. Better measurement is needed to verify the instability of these AI technologies and measure its intelligence capacity: what it can achieve, when and on what it will fail, etc.

Turing Test is subjective [1]. It is an empirical test, not a scientific experiment. Language complexity is much less than the complexity of human intelligence. So Turing Test is invalid.

Sciences are different from mathematics. Scientific experiments can only falsify, but never prove unlimited possibilities. So scientific research should be an ongoing process, always opening to new experiments, especially for immature AI which still does not have scientific foundation.

Existing tests for AI technologies are empirical and inadequate, such as the regular Go games played by AlphaGo Zero and other computer Go systems, the simulations and road tests for self-driving cars, the datasets for natural language understanding, etc.

Driverless cars with no constrains (i.e. SAE level 5 automated-driving) are impossible. There are problems in the definition of SAE level 4, and there is no way to prove a car meet with SAE level 4, especially when the mode evolution in future is not stable.

Technological Singularity is baseless. AI cannot process complicated high-order logic, and cannot recognize complicated sophism, which could cause severe problems in juridical practice, scientific researches, education, medical

practice, etc.

These are very urge problems to study now. In this paper, I will discuss the problems in Turing Test, the problems in current testing methods for AlphoGo Zero, self-driving cars, natural language processing (NLP) [2], and the problems in a mainstream textbook AI: A Modern Approach. Then I will introduce a concept: integrity of intelligence [3], and propose Gu Test: a progressive measurement of generic artificial intelligence based on falsifiability, to solve some critical problems in intelligence studies.

2. The Problems in Turing Test

Turing Test is invalid, and still causes misleading widely in AI research so far. Many current testing methods for AI technologies have similar problems as Turing Test. So it is important to analyze its problems and clarify the misleading.

Turing Test is subjective. People with different knowledge, especially with different levels of understanding about computer and AI technologies, could yield very different results. Based on new scientific progresses, people could find new ways to fail Turing Test even after AI has already passed Turing Test. The subjectiveness of Turing Test causes unstable results. Passing Turing Test does not mean AI has reached or surpassed human level intelligence.

Turing Test is invalid, also because of judging intelligence only based on language conversation. Language complexity is much less than the complexity of human intelligence. Humans have much more intelligence beyond language level intelligence [4]. Indistinguishability between humans and computers by language conversations does not mean equivalence of intelligence.

Turing Test is an empirical test, not a scientific experiment.

Scientific experiments should be done with strictly controlled conditions, to test the underlying principles. Scientific conclusions can only be derived from these principles based on the strict controlled conditions. From empirical tests, people cannot derive scientific conclusions.

Sciences are also different from mathematics. Scientific experiments can only falsify, but never prove unlimited possibilities. Even if AI passes some

experiments stably, people still could find new human intelligence abilities not in AI and design new experiments to fail AI. So, equivalence of intelligence between computers and humans can never be proved, but only can be falsified. Scientific research is an ongoing process, should always open to new experiments.

The other existing testing methods for AI technologies also have many problems. I will discuss the testing problems of computer Go systems, self-driving cars, and NLP later in this paper.

3. AlphaGo Zero's Superhuman Claim

The AlphaGo Zero paper in Nature magazine [5] claimed superhuman performance. However it did not provide any evidences for this claim. Superhuman relates to generic human. The paper did not provide any evidences to show AlphaGo Zero is superior to generic human even in Go gaming.

AlphaGo Zero defeated AlphaGo Master and other computer Go systems is not an evidence of superhuman, because these computer Go systems suffer from the same limitations of AI.

Even if AlphaGo Zero could defeat all human players in regular Go games, this still does not provide evidences for superhuman, because these human players still do not know the limitations of AlphaGo Zero and AI yet.

With scientific experiments, people could discover the weakness of AlphaGo Zero. Then human players could defeat AlphaGo Zero and other computer Go systems in fair Go games later [6].

Scientific experiments are different from regular gaming. Regular gaming is just to win one or some games. Scientific experiments are to falsify some assumptions or let the assumptions pass the experiments. Scientific research should always open to new experiments. In scientific researches, there is no such a thing called "game over".

As said, scientific experiments need strictly controlled conditions. The conclusions of the experiments can only be derived based on the controlled conditions and the results.

Scientific experiment results should open to discussions, so other people

could discuss whether the experiments are valid and whether the interpretations of these results are justified or not.

The Go games played by AlphaGo Zero and other computer Go systems are just regular games or empirical tests, not scientific experiments. People cannot derive scientific conclusions from these games or empirical tests.

So the superhuman claim from the AlphaGo Zero paper in Nature magazine is not a scientific conclusion. As already analyzed, this superhuman claim is baseless.

Go is a game with simple rules and good abstraction, but still with significant complexity. It is much easier to isolate various factors in Go games and figure out some intelligence principles in simple experiments of Go with strictly controlled conditions. These make Go an ideal tool to design scientific experiments on it [7].

4. The Testing of Automated Driving

The current simulations and road tests for self-driving cars are not scientific experiments. They are just empirical tests, without strictly controlled conditions covering all or most cases [8]. So from these simulations and road tests, we cannot derive any scientific conclusions [9].

There is no way to test all or most cases. However, sciences provide methods to approach all cases indirectly. By abstraction, people could gradually discover underlying principles which intend to address all cases. Then people could design experiments to falsify these principles. By abstraction and falsifying principles, people could improve the principles addressing all cases.

Driverless cars with no constrains (i.e. SAE level 5 automated-driving) are impossible, which could be verified with scientific experiments by falsifying the AI technologies used by these driverless cars.

There are problems in the definition of SAE level 4. In different geographical areas, there are very different requirements for SAE level 4 automated-driving techniques. Even in the same areas, if the mode evolution in future is not stable, severe problems could appear with large probabilities even if the cars have already passed SAE level 4 in old modes.

Scientific experiments with the AI technologies used by self-driving cars

could verify that when the mode evolution is not stable, severe problems with large probabilities could occur in the systems which have very few problems in previous modes. So in reality, there is no way to prove a car has truly passed SAE level 4 automated-driving.

Although people could solve some problems in self-driving cars case by case, if they do not know the underlying principles, they would not solve certain important root causes or abstraction of the problems. These root causes could appear in very different forms in future, especially if the mode evolution is not stable [10].

To understand unstable mode evolution, we need understand the underlying principles of intelligence in these AI technologies first. So it is luck that we could study AI problems by studying intelligence principles and designing experiments of these principles with simpler systems, such as AlphaGo Zero.

I wrote before: Go gaming is strictly defined within a very small space. Industrial automation is typically designed in environments well controlled, but not strictly defined. Car driving is regulated, but the environment is not well controlled.

Many industrial automation problems cannot be solved yet. As said, there are still serious problems even in AlphaGo Zero which could be verified by scientific experiments.

Based on my previous analyses and my experiment plans, I have reasons to believe: many years after large-scale deployment of self-driving cars, regular people would have enough chance to interact with self-driving cars from different aspects and trigger unstable mode evolution which causes severe problems with large probabilities. However, at that time, it would be too late.

The technologies for traditional vehicles, such as electronics, powertrain, and other mechanics, etc. are based on concrete sciences whose main principles are already well tested in sciences. However, AI technologies for automated driving are empirical, with no scientific foundation. It could be very unstable in future mode evolution. So testing automated driving vehicles in a similar way as testing traditional vehicles is misleading.

I do not study automated driving directly myself. It is better to do fundamental studies first and figure out the underlying principles of intelligence. By verifying these principles and the problems in AI with simpler and undangerous AI systems, we could avoid the problems of

deploying immature self-driving cars in large scale.

Next, I will discuss the problems in the textbook AI: A Modern Approach. After that, I will discuss language intelligence because it involves more issues than the textbook.

5. The Problems in AI: A Modern Approach

There are problems in the philosophical foundation of the mainstream textbook AI: A Modern Approach, the 3rd edition [11], which directly lead to the problems in testing theories and methods of AI technologies in this textbook.

In the 1st edition of this book, there is an introduction of Socrates, and a reference: "Socrates asks Euthyphro, 'I want to know what is characteristic of piety which makes all actions pious ...'". Socrates was actually talking about a mode of mind, an important topic of intelligence studies. Unfortunately, these contents disappeared in the 3rd edition of this book.

Galileo actually set Socratic method and experiments as the foundation of sciences [12]. By removing the introduction of Socrates, the textbook removed one pillar of sciences. Experiment method, another pillar of sciences, does not exist in the textbook due to its problems of Aristotle thinking mode, the relevance of Turing Test, and Wind tunnel approach. So the textbook does not have scientific foundation. These are also the main problems in current AI sector [13].

I will discuss these problems one by one.

The textbook states: "Aristotle... was the first to formulate a precise set of laws governing the rational part of the mind", which is not true in physical sciences, biology, and mathematics.

The attitude of Aristotle thinking mode is like this: that's it, you have to believe me, there is no need to do experiments (although they are humans, not God). This is exactly what Galileo criticized. Although in different fields, they could appear in very different forms, and derive very different assertions.

As said, the rationale of sciences is based on Socratic method and experiments. Lacking of these elements, Aristotle thinking mode does not have integrity [14] in physical sciences. Actually Galileo and Boyle formally denounced Aristotle thinking mode.

Robert Boyle further suggested the essences of matters rely on their

internal compositions and structures, and should not be confused by their external characteristics as Aristotle thinking mode did [15]. Without understanding the internal structures and complexities of intelligence, AI has the similar problems of Aristotle thinking mode.

Aristotle's biological classification is static, and based on external characteristics, too. Darwin's evolution theory suggested species are dynamic and in evolution, which implies certain internal ambiguity in biological classification. Modern biology tries to improve biological classification by gene and other studies, which actually causes more ambiguity, even spreading to the top level classification. So the ambiguity in biological classification is fundamental, and cannot be eliminated by logic [16]. Aristotle thinking mode does not have integrity in biology.

In mathematics, Gödel even proved the problems of Aristotle's syllogisms. The rationale of mathematics is different from Aristotle thinking mode, and needs more intelligence components and expressing powers, which should be studied further in depth.

So the rationales of physical sciences, biology, mathematics, conflicts with Aristotle thinking mode. The later could not be "a precise set of laws governing the rational part of the mind" [17]. Lacking of integrity of intelligence, Aristotle thinking mode could cause severe problems in many AI applications.

The textbook also states: "Turing deserves credit for designing a test that remains relevant 60 years later", which is obviously not true. As analyzed in section 1, Turing Test causes severe misleading. To understand intelligence and test AI correctly, we need clarify such misleading.

The textbook promotes a Wind Tunnel approach. However, the design of workable wind tunnels, engines, and airplanes all depends on physical sciences. The status of physical sciences in Wright Brothers' age was already very mature, completely different from the status of intelligence studies today which still does not have scientific foundation.

Forces could be understood correctly only after Galileo made critical abstraction over stillness and movement. George Cayley could figure out the underlying principles and forces of flight only after Newton formed the systematic theory of forces. Without these researches, even with a wind tunnel, Aristotle or even Leonardo da Vinci, could not design an airplane successfully.

However, in intelligence studies, we still do not know the fundamental principles, to identify the problems of Aristotle thinking mode and make scientific breakthrough; we still do not know the reasons and complexities

of integrity which is essential to scientific development; we still do not know how to analyze the unstable mode evolution in future which is much more critical in intelligence than in physical sciences.

So Wind Tunnel approach would not work for intelligence studies now. We need structural and systematic analyses of human intelligence first. The studies of language intelligence could provide many important insights.

6. Measure Language Intelligence

AI could do searches well, and have a much better memory for text contents than humans, etc. AI even could achieve many progresses in machine translation. However, AI does not really understand semantics. There is a Chinese room issue, which could be verified.

AI could not process high-order logic properly, could not recognize complicated sophism, could not recognize problems in thinking modes, such as in Aristotle thinking mode or computer thinking mode, etc.

So relying on AI to make judgement could cause severe problems in juridical practice, scientific researches, education, medical practice, etc. Asking students to obey computer thinking mode could damage the development of their intelligence.

The current testing datasets for language understanding, such as the series of SQuAD, CoQA, QuAC, Glue, NLVR^2(Natural Language for Visual Reasoning for Real), cannot measure the exact difference between human and NLP. They cannot help much on high-order logic processing, recognizing sophism, verify Chinese room issues, etc.

All of these datasets fall into the traps of Aristotle thinking mode. They cannot well recognize the problems in thinking modes and the transitions between thinking modes. They cannot identify and verify the principles of intelligence. They are not scientific methods.

To understand human intelligence, we need structural and systematic analyses of human intelligence. I defined certain main intelligence levels: language level, philosophical level, mathematical level, scientific level, with different criteria and requirements, to describe and measure human level intelligence accurately.

Language intelligence is an important characteristic of human intelligence. Other known lives do not have advanced language ability. Language is also

an important media for human knowledge, the basis for philosophy, mathematics, sciences, etc.

Based on languages, humans developed two important branches of studies: mathematics and philosophy. Mathematics develops towards accuracy. Philosophy develops towards integrity.

Sciences originate from philosophy. So sciences also develop towards integrity. Among various alternative theories, only one is correct at most. Beyond philosophy, sciences make conclusions based on experiments by falsifiability with strictly controlled conditions. Sciences also gradually introduce accuracy and mathematics.

So sciences could provide good judgement, which is different from mathematics. Mathematics does not meet the criteria of sciences. It even does not have the property of integrity, does not have a concrete base of experiments [18].

Based on structural and systematic studies of human intelligence [19], people could measure language intelligence much better and much accurately.

7. Integrity and Gu Test

I introduce a new concept: integrity of intelligence, to clarify some misunderstanding and solve certain critical problems in brain and intelligence studies.

Lives at gene or animal level do not have integrity [20]. So there are severe problems in The Selfish Gene (or The Immortal Gene) theory.

Regular natural language expressions or conversations do not have integrity. Debates themselves do not guarantee integrity. Sophism causes severe problems.

Gödel actually proved even mathematics does not have integrity of intelligence. Turing Machine is a subset of mathematics, which also does not have integrity of intelligence, either.

The name "universal approximation theorem" could be misleading, because there is a strict limitation of what could be approximated. Even the simple three-body problem in physics could cause issues. Brain and intelligence

studies are much more complicated. Artificial neural networks with sensors do not have integrity of intelligence.

However, sciences develop with integrity as the goal. The sciences of intelligence need study how integrity and its complexity develop structurally and systematically.

I have been designing a procedure to measure human specific intelligence progressively in AI technologies [21] and discover their problems based on falsifiability, to avoid the problems in empirical tests, and study what conditions could trigger unstable mode evolution and significantly increase the probabilities of these AI problems to dangerous levels.

The test procedure also could provide some insights how integrity develops and how the fundamental principles of intelligence work under strictly controlled conditions.

Based on my studies, I propose:

1) A 4-dimension experiment space to test the intelligence of computer Go systems in Go games and discover the problems in the AI technologies used by these systems, especially to test AlphaGo Zero's superhuman claim or any such implications.

Since there is only one opportunity to gather certain important experiment results before computer Go systems could be adjusted by humans, the first round experiment should be done on the strongest Computer Go system with large-scale experiments [22].

I choice to start experiments with computer Go systems, because some important factors could be well isolated in such simpler systems with good abstraction.

2) Research and experiment schemes for languages, to study the expressing power and limitations of various languages, including natural languages, mathematics, music, etc., and study personalities, thinking modes, and mode transitions behind language expressing, etc.

To understand human intelligence, we need study the development of commonsense, concepts, principles, theories, etc. We need study the fundamental principles in the development of natural languages, philosophy, mathematics, and sciences, especially the development of integrity.

To understand the problems in AI technologies, we need study high-order

logic processing, sophism, Chinese room issues, etc.; and study the problems in computer thinking mode and Aristotle thinking mode, etc.

Existing testing dataset series, such as SQuAD, CoQA, QuAC, GLUE, NLVR^2, etc., do not help much on these issues.

3) Plans to study the relations between brains, mind, and human specific intelligence, started from the problems in The Selfish Gene (or The Immortal Gene) theory. Dream is another important topic.

Experiments could be designed to verify some problems and principles. The experiments will not be related to humans. First, a survey of current brain researches is needed.

Current progresses in neurosciences are mainly at physiological or animal level, such as vision, audio, motion, emotion, etc., which do not illustrate the essentials of human specific intelligence, and cannot show how humans develop integrity based on life entities without integrity.

.....

These studies and experiments require certain amount of resources. They are all fundamental researches for peaceful purposes with no profit prospect. They are scientific researches of the concepts, principles, theories, philosophies of intelligence, rather than personal or social tactics or maneuvers, etc.

Scientific disputes can only be resolved by experiments with strictly controlled conditions. No person could play the role of judge in sciences. Questions and negations to my opinions should be subject to open discussions and strict experiments before conclusions being made.

8. Future Researches

The studies and experiments could be extended in future, to other AI technologies and systems, and to other aspects of human specific intelligence, etc.

More studies should be done to illustrate the reasoning and complexity of integrity development structurally and systematically. However, I need resources to do further researches.

I study these critical problems for the welfare of all humans. However, my health degrades very quickly, some degrading could be irreversible. I cannot do further researches unless in safety personally and economically.

Sophism, misleading, and wrong interpretation of technologies and empirical test results, could damage scientific researches in future.

---

[1] Scientific research should be objective. The scientific principles in quantum physics are still objective, although quantum physics does introduce uncertainty. How to develop objective principles based on the uncertainty in quantum physics is a very important issue of scientific philosophy.

[2] I discuss AlphaGo Zero and self-driving cars, because the superhuman claim of AlphaGo Zero was published in an important academic magazine Nature, and self-driving cars were widely advocated fo many years (called driverless cars before) and relate to public safety. Several years ago, I already heard that the technologies of driverless cars were already ready, just the laws were behind, which obviously is not true.

However natural language is more important, because it is an important media of various human knowledge, and the foundation of philosophy, mathematics, and sciences, etc.

[3] The 'integrity' concept I introduced in intelligence studies is in academic research sense or intelligence sense, not in moral sense.

[4] In section 5., I will discuss more on different intelligence levels.

[5] Mastering the game of Go without human knowledge, published in Nature, on 18 October 2017: https://www.nature.com/articles/nature24270

[6] Actually I designed such experiments as introduced in the section 6 of this article, and requested Deepmind to do the experiments, but they have not accepted the experiments.

Scientific research should be based on open discussion and fair experiment. So the superhuman claim for AlphaGo Zero is not a scientific conclusion.

[7] I had begun to consider Go game as a tool to measure AI technologies long before Deepmind started AlphaGo project.

[8] The current simulations or road tests even could not cover most cases if considering all the possible evolution in future.

An easy way to verify the simulations would not cover all or most cases for self-driving cars is to do experiments with AlphaGo Zero. If AlphaGo Zero could not cover all cases on a 19x19 small board, there is no way to generate simulations covering all or most cases for the much more complicated situations of self-driving cars.

[9] According to some news, in 2015 a blind man was allowed to take a driverless car alone, before the accident on 02/14/2016. Although the damage of this accident is minor, wrong judgment of driverless cars is very dangerous potentially, especially if the mode evolution in future is unstable.

"Steve Mahan, who is legally blind, was the first non-Google employee to ride alone in the company's gumdrop-shaped autonomous car. The ride was in October 2015 in Austin. (Courtesy Waymo)", https://www.washingtonpost.com/local/trafficandcommuting/blind-man-sets-out-alone-in-googles-driverless-car/2016/12/13/f523ef42-c13d-11e6-8422-eac61c0ef74d_story.html,

"Steve Mahan, who is legally blind, takes what Waymo called the world's first fully autonomous ride in Austin in 2015, in an image provided by the Alphabet unit.", https://www.marketwatch.com/story/google-says-driverless-cars-are-ready-to-make-money-but-we-wont-know-if-they-do-2016-12-13.

[10] In a MIT lecture published on Feb 12, 2019, Drago Anguelov. a Principal Scientist at Waymo, admitted that there is a long tail of problems in self-driving cars: https://www.youtube.com/watch?v=Q0nGo2-y0xY.

The real situation could be more complicated than a long tail. By discovering the underlying principles of intelligence we could understand how the problems evolve, and transform, etc.

[11] The 3rd edition of AI: A Modern Approach is referred simply as "the textbook" in this section for convenience.

[12] Dialogue Concerning the Two Chief World Systems, Galileo Galilei (1632).

[13] I inquired Waymo about some recent news of unreasonable behaviors of their cars, and asked them either to deny or to confirm the news reports. I also requested experiments with AlphaGo Zero to verify some generic problems in AI technologies. They have not replied.

Of course, they have the right not to reply. However, by neither denying nor confirming the unreasonable behaviors of their cars, they actually declare they do not follow scientific ways, which raises serious concerns because self-driving cars relate to public safety.

If the news reports are true, there could be serious problems in the cars. Even if they already fixed the problems, the root cause or the abstraction of the problems could still be there. If the reports are not true, not clarifying them also could be dangerous.

Reverse usage of "wolf is coming" is dangerous, too. If faked news of problems (false warning of "wolf is coming") appeared several times without being clarified, then when real problems appears later (wolf is really coming) people would ignore it.

I do not make conclusions based on news or empirical results. However, I do urge Waymo and Deepmind to clarify the issues by open discussion and fair experiments.

Sciences require open discussion and experiments with strictly controlled conditions. Scientific conclusion only could be derived from strictly controlled conditions. Open discussion is to assure the experiments are valid and the interpretation of the experiment results is correct.

[14] Obviously, Aristotle thinking mode does not have integrity even in intelligence sense (not consider moral sense here): Aristotle "counsels Alexander to be 'a leader to the Greeks and a despot to the barbarians, to look after the former as after friends and relatives, and to deal with the latter as with beasts or plants'", Alexander of Macedon, Green, Peter (1991), University of California Press. ISBN 978-0-520-27586-7. https://en.wikipedia.org/wiki/Aristotle.

By dealing other humans as beasts or plants, Aristotle damaged his intelligence and humanity in certain degree. Later, Aristotle had to flee from Athen, to escape possible barbarous persecution from some of Greeks (as "friends and relatives" according to his counsel).

In this paper, I only discuss the problems of Aristotle thinking mode in academic researches, specifically in physics, chemistry, biology, mathematics, intelligence studies. I would not discuss whether Aristotle thinking mode could be useful in some non-academic situations.

[15] The Sceptical Chymist, Robert Boyle (1661).

[16] We need a structural and systematic analyses of intelligence to understand the role of taxonomy. Taxonomy is a language level knowledge, useful in certain degree, but not science in strict sense. It has no integrity of intelligence. People should be very cautious not to derive scientific conclusions from taxonomy.

Further studies are needed to understand what roles could be played and what problems could be caused by Aristotle thinking mode in certain non-academic situations.

[17] Bertrand Russell even wrote: "Ever since the beginning of the

seventeenth century, almost every serious intellectual advance has had to begin with an attack on some Aristotelian doctrine; in logic, this is still true at the present day", A History of Western Philosophy (1945).

The beginning of the seventeenth century, is exactly when scientific revolution started. Since then, civilization experienced a fast development. If open discussion and experiments, the two pillars of sciences, are removed in research now, development could slow down, or even go backwards.

[18] What the differences between sciences and mathematics really indicate, is an important research topic in intelligence studies.

[19] For more details, please see my another article: A Structural and Systematic Analysis of Human Knowledge and Studies. However, I need resources to do further researches.

[20] Gene or regular life entities at animal level do not have integrity, although life sciences have integrity as the goal of development. There are differences between lives and life sciences.

[21] Gu Test does not intend to distinguish humans from humans. It only measures the difference between generic human and machines, or between generic human and other animals. However, it could help humans to avoid the problems of computer thinking mode and Aristotle thinking mode, etc. I treat thinking modes differently from persons or personalities.

[22] To control the experiment conditions strictly, I suggest Deepmind submit all the necessary hardware, software, data files, etc., to Congress Library, so all the versions could be verified before each experiment. Then I could submit my experiment plans.

Other people also could design their experiments based on such a service. Actually I do expect other people could design better experiments after my experiments have been done. Such a public service is better provided by the government.